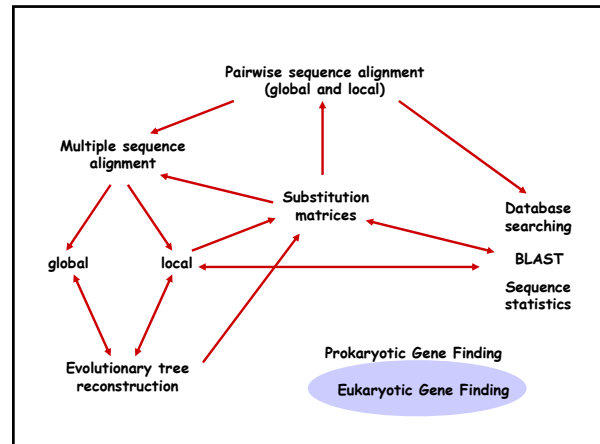


- Tues, Nov 30:  
Gene Finding 1
  - Thurs, Dec 2:  
Gene Finding 2, **PS5 due**
  - Tues, Dec 7:  
Project presentations 1
  - Thurs, Dec 9  
Project presentations 2  
**Final papers due**
  - Tues, Dec 14:  
DD: Extended office hours: 2:30pm – 5:30pm, MI 650
  - Wed, Dec 15  
NS: office hours. DH 1321, noon – 2pm.
  - Friday Dec 17  
**8:30am Final Exam, Room: TBA**
- Online FCE's: Thru Dec 10



## Outline

- **Recap: Prokaryotic gene finding**
- Eukaryotic gene finding
- The human gene complement
- Regulation

## Gene Finding Questions

- Identify protein coding region
- Identify Open Reading Frame
- Predict mRNA (including UTR's)
- Predict intron/exon structure  
Eukaryotes only
- Regulatory signals
- Protein sequence

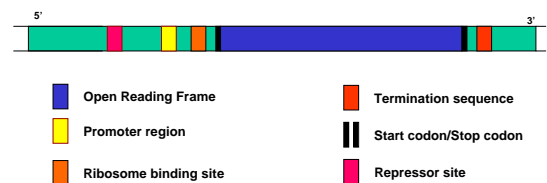
## Gene criteria

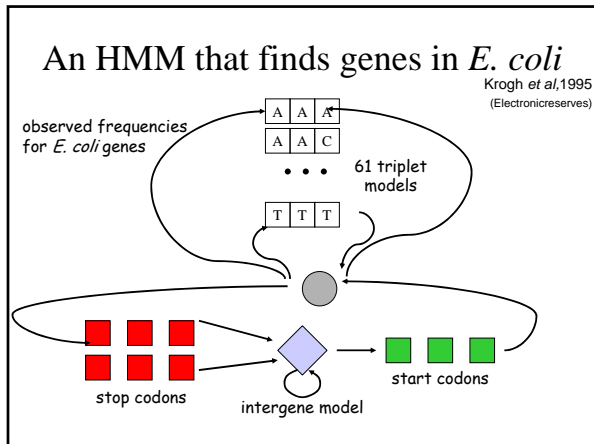
Snyder and Gerstein, Science 2003

- |   |               |
|---|---------------|
| • Open Reading Frames (ORFs)            | Computational |
| • Sequence features                     | Computational |
| • Sequence conservation                 | Computational |
| • Evidence for transcription            | Experimental  |
| • Gene inactivation induces a phenotype | Computational |

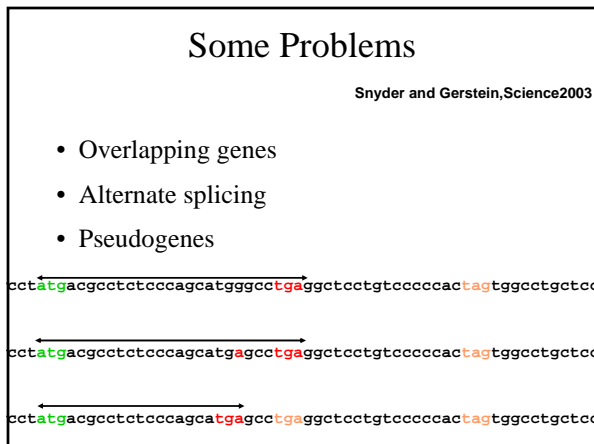
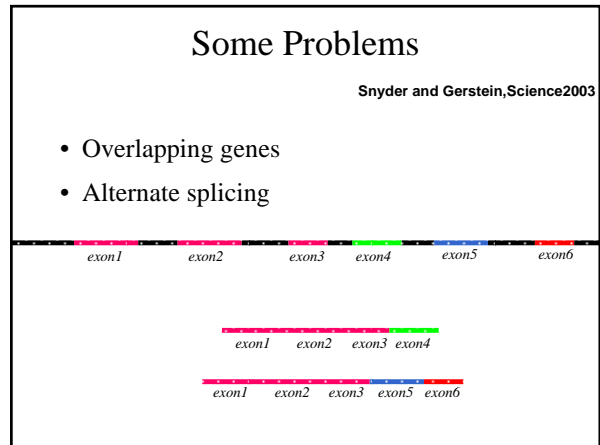
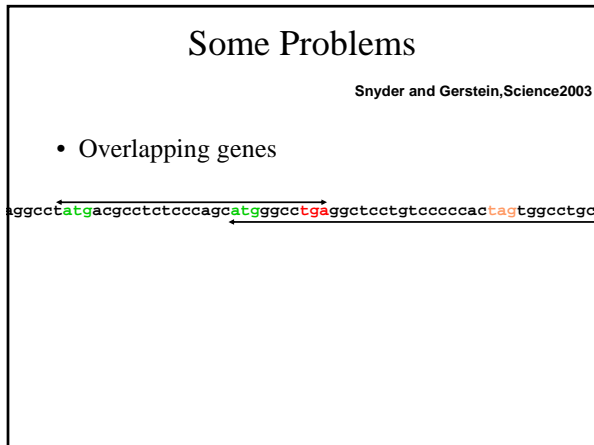
## Sequence features

- Coding statistics (e.g. codon bias)
- Gene structure





- ### Outstanding Problems
- Model cannot account for drift in CG content
  - Does not take position dependencies into account
  - Solution:
    - $k$ th order Markov chain
    - looks back  $k$  positions
  - Glimmer (Salzberg *et al.*, 1998)
    - Finds 98% of all genes in a bacterial genome.



- ### Gene Finding Challenges
- Small protein-coding genes (<100 aa's)
  - RNA-coding genes
  - Regulatory regions
- Salzberg, Nature, 2003
- Genes with sparse conserved positions and little sequence similarity; e.g., beta-defensins
- Schutte *et al.*, PNAS, 2001

## Outline

- Recap: Prokaryotic gene finding
- **Eukaryotic gene finding**
- The human gene complement
- Regulation

## Prokaryotic vs. Eukaryotic Genes

- Prokaryotes
  - small genomes (0.5Mb to 10Mb)
  - high gene density (90%)
  - no introns (or splicing)
  - no RNA processing
  - simple regulatory regions
  - most long ORF's are genes
- Eukaryotes
  - large genomes
  - low gene density (3% - 50%)
  - intron/exon structure
  - splicing
  - complex regulatory regions

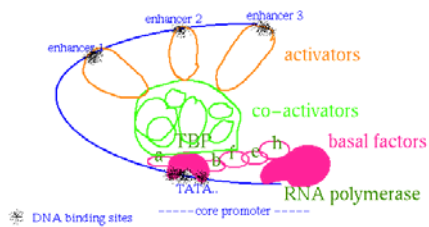
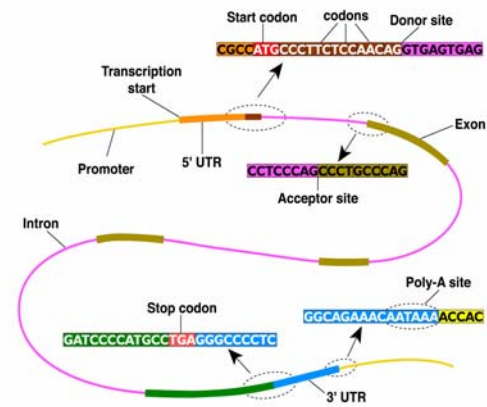
Genomic data:

Must handle multiple genes and/or gene fragments in input sequence.

Typical human gene sizes	
Average gene length	30kb
Coding region	1-2kb
Exon length	150 - 200 bp
Exon count	5-6
Single exon genes	8%

	Genome statistics		
	Size	Gene number	Density (1 gene per)
Human	3300Mb	30K	100,000
Fly	180Mb	13.6K	9000
<i>C. elegans</i>	97Mb	19.1K	5000
Yeast	12Mb	6.3K	2000
<i>E. coli</i>	4.8Mb	3.2K	1400
<i>H. influenzae</i>	1.8Mb	1.7K	1000

[http://www.ornl.gov/TechResources/Human\\_Genome/faq/compngen.html](http://www.ornl.gov/TechResources/Human_Genome/faq/compngen.html)



Source: <http://www.nslj-genetics.org/gene>

## Genscan

Burge and Karlin, 1997

### Architecture:

- Individual modules: intergenic region, promoter, 5' UTR, exon/intron, post-translation region
- Semi Hidden Markov Model – various length distributions
- Different statistical models for each module:
  - weight matrices + extensions, 3-periodic 5<sup>th</sup> order Markov chains

### Incorporates:

- Descriptions of transcriptional, translational and splicing signals
- Compositional features of exons, introns, intergenic, C+G regions

## GenScan

Burge and Karlin, 1997

Larger predictive scope than previous models

- Partial genes
- Multiple genes separated by intergenic DNA
- Genes on either/both DNA strands

Proposed pipeline

- Screen for repetitive elements
- Predict protein sequences with GENSCAN
- BLAST predictions to find homologs
- Refine using spliced alignment of prediction with homolog (e.g., Gelfand, Mironov, Pevzner, 96)
- Verify experimentally

## GenScan States

- N: intergenic region
- P: promoter
- F: 5' untranslated region
- $E_{\text{singl}}$ : single exon (intronless) (translation start  $\rightarrow$  stop codon)
- $E_{\text{init}}$ : initial exon (translation start  $\rightarrow$  donor splice site)
- $E_k$ : phase k internal exon (acceptor splice site  $\rightarrow$  donor splice site)
- $E_{\text{term}}$ : terminal exon (acceptor splice site  $\rightarrow$  stop codon)
- $I_k$ : phase k intron:

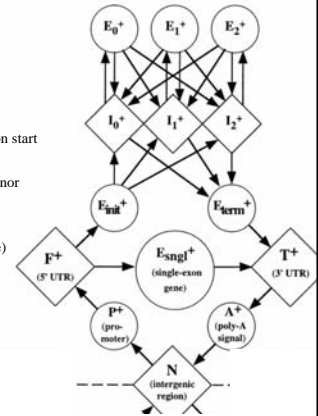


Fig. 3, Burge and Karlin 1997

## GenScan States

- N: intergenic region
- P: promoter
- F: 5' untranslated region
- $E_{\text{singl}}$ : single exon (intronless) (translation start  $\rightarrow$  stop codon)
- $E_{\text{init}}$ : initial exon (translation start  $\rightarrow$  donor splice site)
- $E_k$ : phase k internal exon (acceptor splice site  $\rightarrow$  donor splice site)
- $E_{\text{term}}$ : terminal exon (acceptor splice site  $\rightarrow$  stop codon)
- $I_k$ : phase k intron:

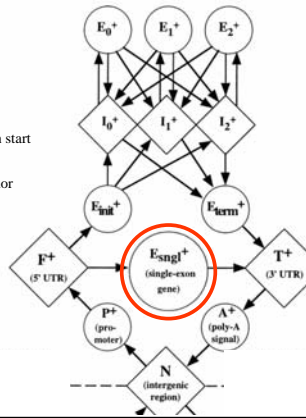
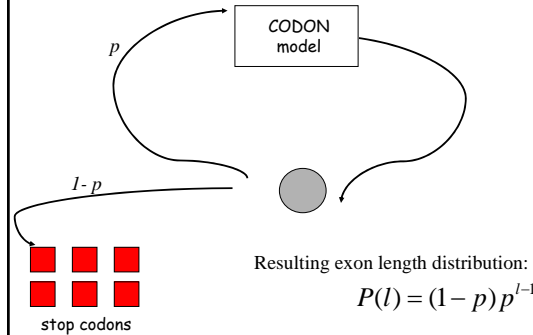


Fig. 3, Burge and Karlin 1997

How to model sequences with lengths that are not geometrically distributed?



## Semi-hidden Markov model

- Set of states:  $Q1, Q2, \dots$
- Transition matrix  $P(Q(t)|Q(t-1))$
- Initial distribution  $P(Q_i)$
- Each state has
  - a length distribution
  - a sequence generating model
- Emission:
  - Each state emits a *sequence*, according to a particular distribution, of length,  $d$ , according to a particular length frequency distribution

## Semi-hidden Markov model cont'd

- A parse  $\varphi$  of length  $L$  is
  - A state sequence:  $Q1, Q2, \dots$
  - A sequence of lengths:  $d1, d2, d3, \dots$
- An observed sequence,  $s$ , is scored using a modified Viterbi algorithm

$$\varphi_{\text{opt}} = \arg \max P(\varphi | s)$$

## GenScan Training Set

2.5M base pairs

142 Single Exon Genes (SEGs)

238 multi-exon gene

1492 Exons

1254 Introns

An additional 1619 coding sequences (no introns)

Promotor model based on published sources.

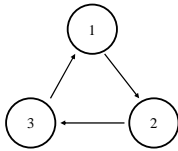
## Initial and transition probabilities

Trained separately for four categories of G+C content

- < 43% (G+C)
- 43% - 51% (G+C)
- 51% - 57% (G+C)
- > 57% (G+C)

- Gene density varies with G+C content
- Genes in A+T rich regions had fewer introns

## Single Gene Exons



- 3-periodic
- 5<sup>th</sup>-order Markov chain
- Length distribution:  
taken empirically from single  
exon lengths
- Sequence model  
Trained with coding sequences

## GenScan States

- N: intergenic region
- P: promoter
- F: 5' untranslated region
- E<sub>sngl</sub>: single exon (intronless) (translation start -> stop codon)
- E<sub>init</sub>: initial exon (translation start -> donor splice site)
- E<sub>k</sub>: phase k internal exon (acceptor splice site -> donor splice site)
- E<sub>term</sub>: terminal exon (acceptor splice site -> stop codon)
- I<sub>k</sub>: phase k intron:

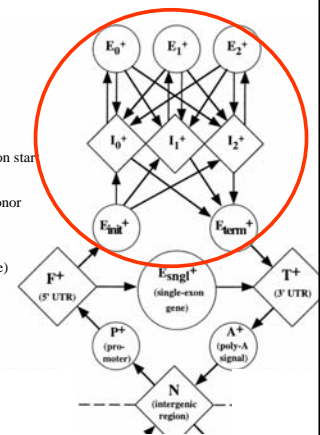


Fig. 3, Burge and Karlin 1997

## Intron/Exon model

phase 0 intron:

GAGCATGACTxxxxxxxxxxxxxxxxGAGGTGCACCTGGTGACTTAGAC..

phase 1 intron:

GAGCATGACTGxxxxxxxxxxxxxxxxAGGTGCACCTGGTGACTTAGAC..

phase 2 intron:

GAGCATGACTGxxxxxxxxxxxxxxxxGGTGCACCTGGTGACTTAGAC..

- Exon phase = phase of previous intron
- Donor and acceptor models incorporated in intron models
- Length and sequence distribution determined empirically

## GenScan States

- N: intergenic region
- P: promoter
- F: 5' untranslated region
- E<sub>sngl</sub>: single exon (intronless) (translation start -> stop codon)
- E<sub>init</sub>: initial exon (translation start -> donor splice site)
- E<sub>k</sub>: phase k internal exon (acceptor splice site -> donor splice site)
- E<sub>term</sub>: terminal exon (acceptor splice site -> stop codon)
- I<sub>k</sub>: phase k intron:

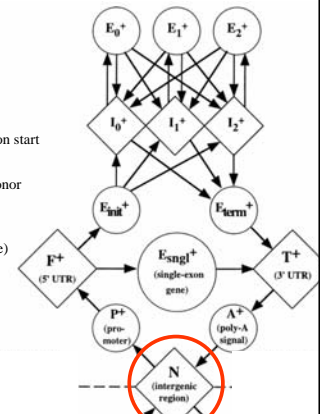


Fig. 3, Burge and Karlin 1997

## Intergenic models

- Lengths are geometrically distributed

$$\text{mean: } \frac{3 \times 10^9}{60,000} \quad (\text{estimated human gene number in 1997})$$

- Sequence model
  - 5<sup>th</sup> order Markov model (*highest order trainable with data available*).
- Similar models used for untranslated regions

## GenScan States

- N: intergenic region
- P: promoter
- F: 5' untranslated region
- E<sub>sngl</sub>: single exon (intronless) (translation start -> stop codon)
- E<sub>init</sub>: initial exon (translation start -> donor splice site)
- E<sub>k</sub>: phase k internal exon (acceptor splice site -> donor splice site)
- E<sub>term</sub>: terminal exon (acceptor splice site -> stop codon)
- I<sub>k</sub>: phase k intron:

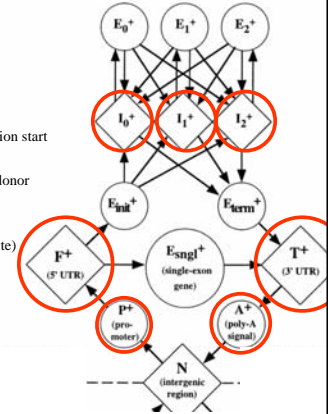
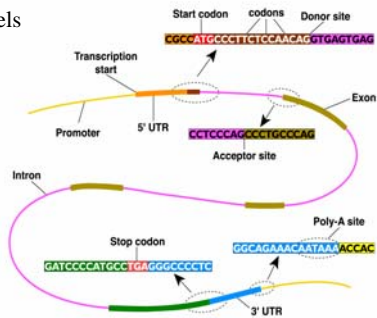


Fig. 3, Burge and Karlin 1997

## Signal Models

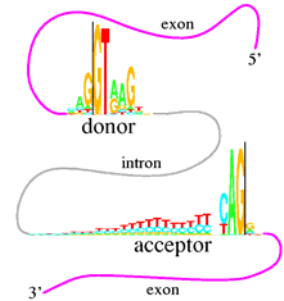
TATA box, polyA signal, 5' UTR

- Fixed length models
- PSSMs
- No position dependence



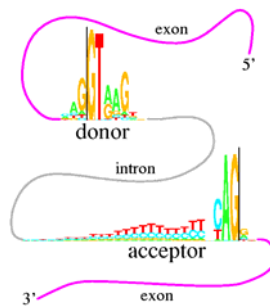
## Acceptor Model

- Fixed length model
- “Weight array model”
  - Models dependence on the previous position



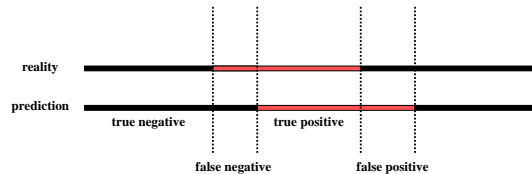
## Donor Model

- Fixed length model
- “Maximal Dependence Decomposition”
  - Models dependencies between nonadjacent elements



## Performance measures

Burset & Guigo, 1996

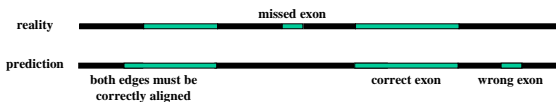


Nucleotide Level

$$S_n = \frac{TP}{TP + FN}$$

## Performance measures

Burset & Guigo, 1996



Exon Level

$$S_n = \frac{TP}{TP + FN}$$

$$S_p = \frac{TN}{TN + FP}$$

## Genscan Performance

	Nucl Sn	Exon Sn	Exon Sp	Missing Exons	Wrong Exons
Genscan	0.93	0.78	0.81	0.09	0.05
FGENEH	0.77	0.61	0.64	0.15	0.12
GENEID+	0.91	0.73	0.70	0.07	0.13
GENEPARSER3	0.86	0.56	0.58	0.14	0.09

Proportion of genes with all exons correctly predicted:

$$\frac{243}{570} = 0.43$$

## Outline

- Recap: Prokaryotic gene finding
- Eukaryotic gene finding
- **The human gene complement**
- Regulation

## Gene Prediction in the Human Genome

International Human Genome Consortium, *Nature* 2001

Initial Gene Index (IGI)

- Ensembl
  - GENSCAN predictions confirmed by homology to EST's, mRNAs, proteins and protein motifs
- Genie
  - HMM-based
  - Extends homology with EST's/mRNAs using *ab initio* approach
- Previously known genes in data bases

14882	Known genes
4057	Ensembl + Genie
12,839	Ensembl alone
31,778	Total Predictions

## Validation

- Comparison with 31 newly discovered genes
  - 28/31 in draft human genome sequence
  - 19/28 were detected by gene prediction
  - IGI contains ~60% of novel genes in human genome (19/31)
- Comparison with mouse cDNA's
  - 81% of mouse cDNA's similar to draft human sequence
  - 69% of mouse cDNA's similar to predicted genes
- Problems:
  - Overprediction
  - Fragmentation: 1 true gene corresponds to >1 prediction
  - Partial prediction: only part of gene is correctly predicted

## Predicting Human Genome Count

IGI contains:

15,000 known genes  
17,000 predicted genes

yields 32,000 genes

Assuming

20% overprediction  
fragmentation rate of 1.4

yields 24,500 genes

Assuming

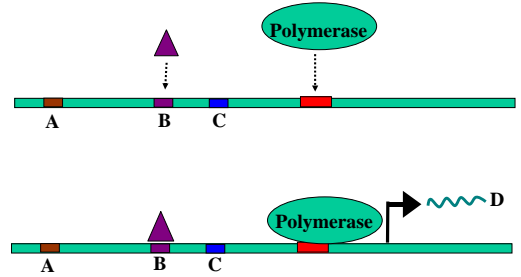
40% of novel genes are not predicted

yields **31,000 genes**

## Outline

- Recap: Prokaryotic gene finding
- Eukaryotic gene finding
- The human gene complement
- Regulation

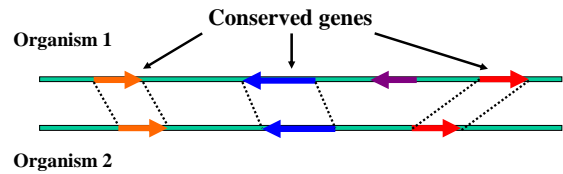
## Regulatory regions



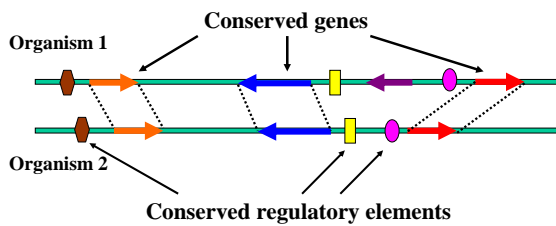
## Examples of binding site motifs

<u>A</u> CC <u>ACA</u>	Rap1
SGTGGCAAA	Rpo4
GAATCA	Gcn4
CTGAAITC	HSE
TCC	Mig1/STRE
CCAATA	Hap2,3,4
CACGTGA	Clb1
ACGCGT	MCB
TTCGAAAT	Lys14
CCGT	Leu3

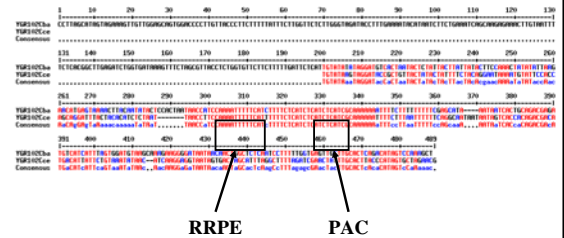
## Identifying binding sites Comparative genomic approach



## Identifying binding sites Comparative genomic approach

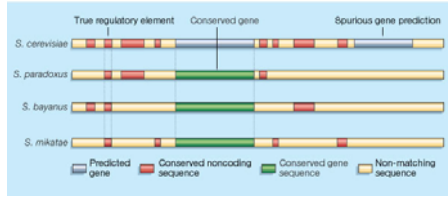


## Global alignment of upstream sequences



Kellis *et al.*, Nature, 2003

## Comparative genomics for gene finding



Salzberg, Nature, 2003

## Comparative genomics for gene finding

Using genomic sequence from  
*S. paradoxus*, *S. bayanus*, *S. mikatae*

Kellis *et al.* (Nature, 2003) found

503 false predictions

43 new small genes

42 new regulatory motifs

Requires sequences in species that are close but not too close.

Will this approach work in higher eukaryotes?

One of the first examples of this approach:

***Escherichia coli* promoter sequences predict in vitro RNA polymerase selectivity.** Mulligan ME, Hawley DK, Entriken R, McClure WR. Nucleic Acids Res. 1984 Jan 11;12(1 Pt 2):789-800.

We describe a simple algorithm for computing a homology score for *Escherichia coli* promoters based on DNA sequence alone. The homology score was related to 31 values, measured in vitro, of RNA polymerase selectivity, which we define as the product KBk2, the apparent second order rate constant for open complex formation. We found that promoter strength could be predicted to within a factor of +/-4.1 in KBk2 over a range of 10(4) in the same parameter. The quantitative evaluation was linked to an automated (Apple II) procedure for searching and evaluating possible promoters in DNA sequence files.

	Whole genome sequencing	Gene finding
1992		Assessment of protein coding measures (Fickett & Tung)
1995	<i>H. influenzae</i> , 1 <sup>st</sup> whole genome sequence	Sequence features in coding, non-coding and intergenic DNA (Guigo & Fickett) HMM gene finder for <i>E. Coli</i> (Krogh <i>et al</i> )
1997	Yeast, 1 <sup>st</sup> eukaryote	
1998	<i>C. elegans</i> 1 <sup>st</sup> multicellular organism	<b>Glimmer</b> , Higher order Markov models for prokaryotes (Salzberg <i>et al</i> ) <b>Genscan</b> , Prediction of complete gene structures in human genomic DNA (Burge & Karlin)
2000	Fly, <i>Arabidopsis thaliana</i> (mustard weed) 1 <sup>st</sup> plant	
2001	Human, 1 <sup>st</sup> mammal	
2002	Mouse	Kellis <i>et al</i> , 2003, comparative genomics for finding regulatory regions

